Application of data mining in medical domain: case of cardiology sickness level

Saraee, MH, Waheed, A, Javed, S and Nigam, J

Application of Data Mining in Medical Domain: Case of Cardiology Sickness Level

Mohamad Saraee, Asad Waheed, Sargheel Javed and Joshi Nigam
School of Computing, Science and Engineering
University of Salford

ABSTRACT

Accuracy plays a vital role in the medical field of Cardiology as it concerns with the life of an individual. It is very important and crucial while taking decisions, which includes both the past experience and the present situations. Data mining in the medical domain works on the past experiences (data collected) and analyse them to identify the general trends and probable solutions to the present situations. This paper is concerned with the application of data mining techniques in the domain of the medical field of heart diseases/attack. We carried out extensive experiments applying different data mining techniques including Relevance analysis, Association Rules Mining and Clustering. We report the findings which are very promising.

1. INTRODUCTION

Knowledge can be identified as information associated with rules which allow interferences to be drawn automatically so that information can be used for purposes. In a medical field previous knowledge is used to predict and diagnose for a new condition i.e. if a patient suffered from conditions/symptoms that are very similar to a previous patients the same solution could be deployed.

Data mining is currently implemented in many clinical environments where relationships and patterns provide new medical knowledge. Most of these applications consist of database which holds vast amounts of data. 'The typical data mining processes involves transferring data originally collected in production systems in a data warehouse cleaning or scrubbing the data to remove errors and check for consistency or formats, and then mining the data using the different techniques[1]. The deficiency that exists at present is that various data mining techniques are available but patterns and rules discovered are vary from one another which could resulted in different diagnoses.

In this work Detrano published dataset were selected to be utilised [3]. This dataset consists of information of 303 patients which 165 are free of heart disease and the remaining 138 are with heart attack. Number of data mining algorithms were applied to this dataset to identify the general trends and asses the several outputs. The problem in the area of Cardiology is to predict the sickness level on the basis of the relevant health attribute. we collected the dataset that consists of 303 instances. The different types of data analysis performed consist of Relevance Analysis, Clustering, and Association Rules Mining.

2. METHODOLOGY

This experiment includes converting the extracted medical data into cleaned and arranged in a suitable format which is compatible with the Envisioner data mining tool [1] which we use to conduct the experiment. Relevance analysis will be carried to retrieve the most relevant attributes. Using those attributes the above tools the dataset will be mined. On the basis of relevance analysis performed by Envisioner tool the most relevant attributes that were identified with respect to 'Status' (healthy or Sick) were used as parameters for the classification which can predict the sickness level and status of each patient Relevance analysis was carried out to determine the most relevant attributes to 'Status' and 'Result' with max field constraints up to 4 fields. The results shown in Figure 1 and Figure 2 respectively.

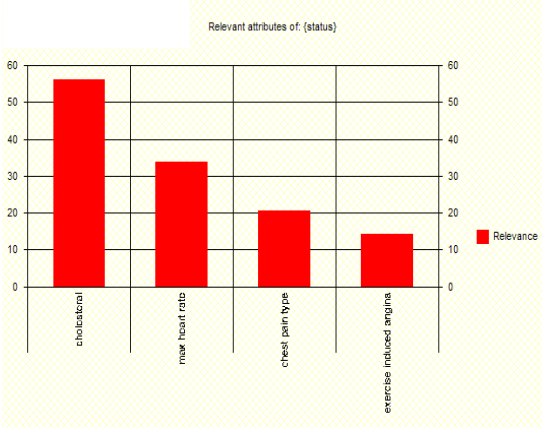


Figure 1: Relevant Analysis (status)

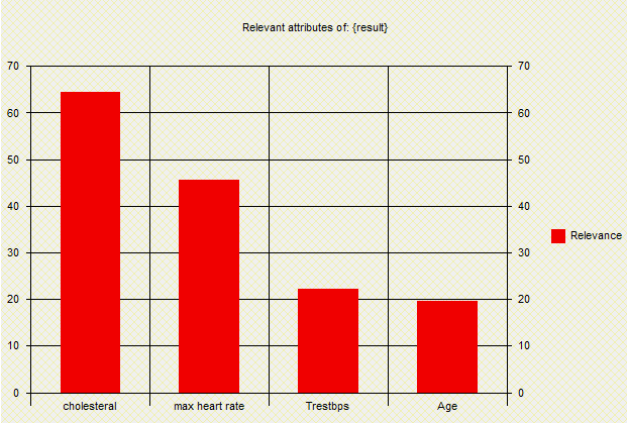


Figure 2: Relevant Analysis (Result)

Most relevant fields are cholesterol, max heart rate. With respect to the 'Result' we get a different relevant attribute called 'Trestbps', which is not in the 'Status' relevance graph. The same analysis is performed using 8 fields. The result is shown in Figure 3.

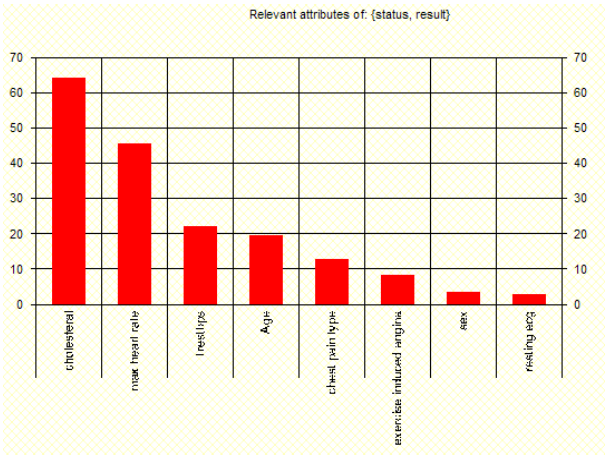


Figure 3: Relevant Analysis (Result)

Analysing the graph of Figure 3 it can be stated that age and sex are less relevant attributes for the occurrence of the heart attack. Analyzing all the three graphs, we can clearly identify that from the 8 attributes only two of which are most relevant to 'Status' & 'Result'. Here is the ascending pattern of attribute relevance which is stated below.

Cholesterol	approx 65%
Max Heart Rate	approx 45%
Chest Pain Type	approx 15%

with "Predicted Result" having probabilities of each attributes.

2.1 Clustering

Clustering is an important exploratory data analysis task. It aims to find the intrinsic structure of data by organizing data objects into similarity groups or clusters. It is often called unsupervised learning because no class labels denoting an *a priori* partition of the objects are given.[2]. Using the Weka application we applied two Clustering algorithms, 'Farthest First' and 'Density Based'.

Density Based Cluster Percentage Split Model: Only the Main points have been displayed. Below is the Outputs of the Clustering procedure.

```
Test mode:    split 66% train, remainder test
=== Clustering model (full training set) ===
MakeDensityBasedClusterer:
Wrapped clusterer:
kMean
Number of iterations: 4
Within cluster sum of squared errors:
109.17519305741266
Cluster centroids:
Cluster 0
            Mean/Mode:   56.661  136.8814 242.1017
131.5424 true S1
            Std Devs:    7.8732  18.8626  49.9098
18.2599 N/A     N/A
Cluster 1
            Mean/Mode:   53.2357 130.2286 245.6786
156.05   fal H
            Std Devs:    9.4135  16.032   55.0021
20.4216 N/A     N/A
Clustered Instances

0       32 ( 31%)
1       72 ( 69%)
Log likelihood: -19.55389
```

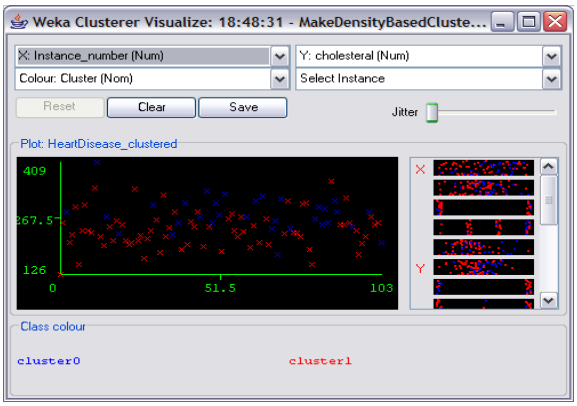


Figure 4: Output Density Based Cluster Percentage Split Model

Farthest First Cluster Percentage Split Model

Observation:*Only the Main points have been displayed

```
Test mode:    split 66% train, remainder test
=== Clustering model (full training set) ===
 FarthestFirst
==============
Cluster centroids:
Cluster 0
            52.0 128.0 204.0 156.0 true S2
Cluster 1
            67.0 115.0 564.0 160.0 fal H
=== Model and evaluation on test split FarthestFirst
Cluster centroids:
Cluster 0
            55.0 132.0 353.0 132.0 true S3
Cluster 1
            29.0 130.0 204.0 202.0 fal H
Clustered Instances
0       40 ( 38%)
1       64 ( 62%)
```

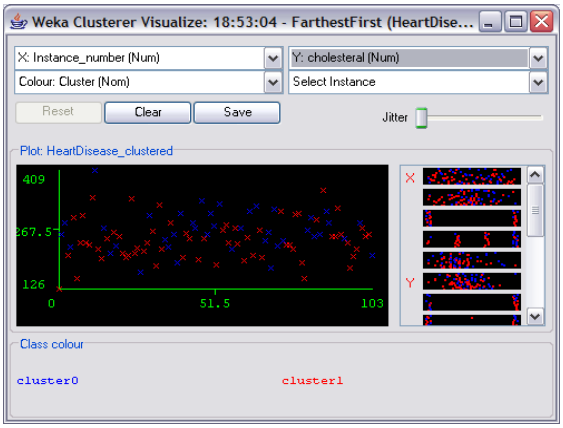


Figure 5: Farthest First Cluster Percentage Split Model

2.2 Association Rules Mining

Association mining on a set of data looks for values in different attributes that commonly occur together, suggesting an Association between them. We have applied Association to the data set on the bases of the relevance analysis carried for the 4 most relevant attributes. We discovered that if cholesterol is in the range of 174 to 245 then status is equal to Healthy/sick but it differs in confidence though the support is the same.

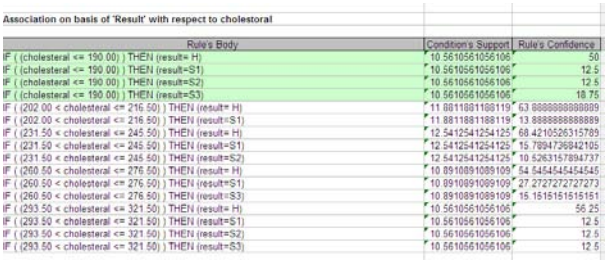
Association on the bases of Status with regard to cholesterol

Example of rules identified:

- 174< cholesterol <=245 :
  Healthy: Support = 10.23%
  Confidence =67.74%
  Sick: Support = 10.23%
  Confidence =32.26%

Association on the bases of Result with regard to cholesterol

The support and the confidence for the attribute result {S1, S2, S3, H} with regard to different ranges of the cholesterol are presented in Figure 9.



Association on basis of 'Result' with respect to cholestoral

Rule's Body	Condition's Support	Rule's Confidence
IF ((cholesteral <= 190.00)) THEN (result= H)	10.5610561056106	50
IF ((cholesteral <= 190.00)) THEN (result=S1)	10.5610561056106	12.5
IF ((cholesteral <= 190.00)) THEN (result=S2)	10.5610561056106	12.5
IF ((cholesteral <= 190.00)) THEN (result=S3)	10.5610561056106	18.75
IF ((202.00 < cholesteral <= 216.50)) THEN (result= H)	11.881188118811 9	63.8888888888889
IF ((202.00 < cholesteral <= 216.50)) THEN (result=S1)	11.881188118811 9	13.8888888888889
IF ((231.50 < cholesteral <= 245.50)) THEN (result= H)	12.541254125412 5	68.4210526315789
IF ((231.50 < cholesteral <= 245.50)) THEN (result=S1)	12.541254125412 5	15.7894736842105
IF ((231.50 < cholesteral <= 245.50)) THEN (result=S2)	12.541254125412 5	10.5263157894737
IF ((260.50 < cholesteral <= 276.50)) THEN (result= H)	10.891089108910 9	54.5454545454545
IF ((260.50 < cholesteral <= 276.50)) THEN (result=S1)	10.891089108910 9	27.2727272727273
IF ((260.50 < cholesteral <= 276.50)) THEN (result=S3)	10.891089108910 9	15.1515151515151
IF ((293.50 < cholesteral <= 321.50)) THEN (result= H)	10.5610561056106	56.25
IF ((293.50 < cholesteral <= 321.50)) THEN (result=S1)	10.5610561056106	12.5
IF ((293.50 < cholesteral <= 321.50)) THEN (result=S2)	10.5610561056106	12.5
IF ((293.50 < cholesteral <= 321.50)) THEN (result=S3)	10.5610561056106	12.5

Figure 6: Support and the confidence for attribute Result with regard to Cholesterol.

Example of characteristics identified:

- cholesterol <=190:
  Support for 'Result' {S1, S2, S3, H} = **10.56%**
  Confidence for 'Result' {S1, S2, S3, H} = {12.5, 12.5, 18.75, 50} respectively.

Association on the bases of Result with regard to Max Heart Rate

The support and the confidence for the attribute result {S1, S2, S3, H} with regard to the different ranges of the cholesterol which presented in Figure 10.

Example of rules identified:

- cholesterol <=190:
  Support for 'Result' {S1, S2, S3, H} = 10.56%
  Confidence for 'Result' {S1, S2, S3, H} = {12.5, 12.5, 18.75, 50} respectively.

For the same result {S1, S2, S3, H} we obtained different ranges of 'Max Heart Rate' with equal support but different Confidence percentage.

4.3 DISCUSSION

On the basis of guidelines given in the data set, status indicates if the patient is either healthy or sick where as the result indicates the actual sickness level if a patient is sick(S1,S2, S3) or "H" for healthy. After carrying out the first part of experiment, which was based on relevance analysis, we identified 4 relevant attributes linked to Status and Result as shown in Figure 1,2 and 3. From the outputs created we can observe a difference between relevant attributes. In Figure 1 which is based on status the attribute Chest pain type and Angina are different than Trestbps and Age in Figure 2. We can identify that Age is more relevant than exercise induced angina having approx. 20% over approx. 15%. Two Clustering algorithms 'Farthest First' and 'Density Based' were applied to the dataset. Weka density based clustering technique implements following models and evaluation on training set:

- Make Density Based Clusters
- Wrapped Clusters
- kMeans

Where as a Farthest First model is implemented for evaluating the training set. We can articulate that while implementing Density Based Clustering model kMeans algorithm is being implemented by default. The final experiment was based on Mining Association Rules. We encountered some unspectacular patterns within the output of 'Association on the basis of Result with regard to Cholesterol' as shown in Figure 6 and 'Association on the basis of Result with regard to Max Heart Rate'. We can see that in any range all the sickness levels (S1, S2, S3, S4, H) has equal support but different confidence. Particularly as shown in Figure 9 in a specific range of cholesterol level. We figured out same confidence for S1 and S2 sickness levels & H (healthy) has maximum confidence of 50%. From such observation it can be concluded that although cholesterol level being maximum relevant attribute it fails to suggest the exact sickness level of the patient.

5. CONCLUSION

Accuracy plays a vital role in the medical field of Cardiology as it concerns with the life of an individual. Henceforth it's very important and crucial while taking decisions, which includes both the past experience and the present situations. Data mining in the medical domain works on the past experiences (data collected) and analyse them to identify the general trends and probable solutions to the present situations.

Henceforth keeping in mind the criticality of the field we choose it for experimenting the various mining techniques and finally we came to a conclusion that taking into a real life scenario one can't predict the exact solution or the action to be taken for the diagnoses of the patient on the basis of machine generated predicted values from the past results. Since every things works on the probability we should even consider the mined results to be probable and not the final.

6. REFERENCES

[1]Neurosoft Envisioner Help - Manual (1999 2000) Version 1.0

[2] Jonathan C. Prather, M.S. "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse". 1995

[3] Dr Dettano Dataset: www1.ics.uni.edu/pub/machine-learning-database/heart-disease/cleve.mod